AI and Ethics (2026) 6:197
https://doi.org/10.1007/s43681-026-01036-z

REVIEW

The emerging relevance of relational justice within algorithmic fairness research: a systematic literature review

Laila Wegner[1] · **Marie Mirsch**[2] · **Carmen Leicht-Scholten**[2]

Received: 19 September 2025 / Accepted: 5 February 2026
© The Author(s) 2026

Abstract
Algorithmically reinforced inequalities recently motivated a growing area of algorithmic fairness research, particularly focusing on statistical metrics to mitigate technical biases. Most of these approaches are rooted in distributive fairness ideals and focus on algorithmic outcomes. However, just as the broader research discipline of AI ethics, a recently emerging third wave signals a new research paradigm within algorithmic fairness research. This shift is conceptualized as a structural turn that also centers power dynamics and structural injustices. Given the dominating topics, we hypothesize that the recent developments within the algorithmic fairness research are best captured by relational accounts of justice. Relational justice extends beyond the distributive ideals and foregrounds social relationships when assessing justice, which includes the rejection of oppression and domination, but treats all individuals as equals based on mutual respect. Therefore, we investigate what the lens of relational justice can add to the (so far merely distributive) paradigm within algorithmic fairness research. Using a systematic literature review, we synthesized insights from 45 scientific contributions that, while potentially not explicitly framed within relational justice, implicitly engage with a relational lens. We categorized themes into eight subject areas. Our analysis offers a novel review of algorithmic fairness research that concentrates on the underlying philosophical discourses, guided by a relational perspective, and thereby embedding relational justice within the field of algorithmic fairness.

Keywords Algorithmic fairness · Relational justice · Structural injustice · Systematic literature review

1 Introduction

Addressing the urgent need to discuss the societal impacts of Artificial Intelligence (AI), particularly algorithmic decision making (ADM), research on algorithmic fairness and justice[1] is constantly growing. However, the research field is marked by significant conceptual ambiguities. Several years ago, it was shown that it is mathematically impossible to satisfy all existing algorithmic fairness metrics simultaneously [1, 2]. Since then, several authors have argued that this focus misses the core issue since algorithmic fairness metrics themselves are misleading or illusory [e.g., 3]. To examine these tensions and the differing perspectives on algorithmic fairness, it is useful to situate the field within its broader research context: AI ethics.

The field of AI ethics has emerged as a critical, interdisciplinary research domain. The term 'artificial intelligence' itself is an equivocal term with many different associated meanings, lacking a clear definition. From a more technical point of view, AI is often used in reference to machine learning, i.e., statistical methods that learn patterns from data to

[1] While fairness and justice in philosophy are distinct concepts, in the field of algorithmic fairness and justice, the distinction is often not clearly defined. For a detailed discussion, see [13]. For the sake of this paper, we associate algorithmic fairness with formal approaches, while the reference to justice indicates a broader societal assessment. Furthermore, the terms used in the presentation of the literature reviews results are shaped by the respective contributions.

✉ Laila Wegner
l.wegner@tue.nl

✉ Marie Mirsch
marie.mirsch@gdi.rwth-aachen.de

Carmen Leicht-Scholten
carmen.leicht@gdi.rwth-aachen.de

[1] Philosophy & Ethics, Eindhoven University of Technology, Eindhoven, Netherlands

[2] Gender and Diversity in Engineering, RWTH Aachen University, Aachen, Germany

Published online: 15 March 2026

make predictions without being explicitly programmed. In ethics and philosophy, AI is more broadly defined as a sociotechnical system, meaning that it is shaped by the interplay between technical components, human behavior, social and organizational practices, and political infrastructures [e.g., 4]. Despite different topics within AI ethics, the research paradigms can be summarized within three 'waves of AI ethics' [5–7]. The first wave was dominated by abstract and high-level guidelines and principles for the development and use of AI systems [6, 7]; or by a focus on long-term issues, dominated by negative and abstract concerns of dystopian technology [5]. The second wave is conceptualized as concrete and technological approaches to formalize the ethical principles in practice, aiming to overcome the abstract and high-level character of guidelines [5–7]. This includes, for example, formal fairness notions to mitigate bias mathematically [8, 9] or to increase explainability [10, 11]. The recently emerging third wave emphasizes that algorithms must be understood as sociotechnical systems, meaning that technology shapes society and vice versa [5–7]. Bolte and van Wynsberghe [5] conceptualize the shift from the second to the third wave as a turn "from an isolationist to a structural approach" (p. 4) that also considers the underlying power structures.

The field of algorithmic fairness and justice, understood as a sub-branch of AI ethics, serves as a representative example for these developments [5]. In the first wave, guidelines such as the European ethics guidelines for trustworthy AI [12] demand "ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatization" (p.12). What exactly is considered just distribution or unfair biases, and how to mitigate them, remains vague in these high-level guidelines. Putting this into practice was the objective of the second wave, as described above [5]. Thus, as part of the second wave, technical research on algorithmic fairness focuses on identifying, measuring, and mitigating discriminatory biases in algorithmic predictions and representations. The discourse commonly centers around legally protected categories [13, 14] and was stimulated by real-world examples such as gender-biased hiring algorithms [15] or race-biased crime sentencing algorithms [16]. Focusing on the distribution of outcomes (often represented by true/false positive/negative predications), computational fairness metrics are typically, at least loosely[2], inspired by distributive ideals of justice [19].

In contrast, then, the third wave of algorithmic fairness indicates a shift in both the main objectives of research and its theoretical grounding. It is dominated by critical contributions that highlight methodological limitations of algorithmic fairness, such as its tendency to focus on technical fixes (often referred to as 'technosolutionism') based on a narrow dimension of bias, rather than on systems of oppression [20]. Instead, research that is attributed to this third wave centers on questions around power, structural inequalities, or sociotechnical analyses. Similarly, philosophical theories of relational justice focus on unequal power asymmetries, social relations, and structural injustices by foregrounding equality based on the quality of social relations and social institutions [21]. Instead of criticizing inequality on the grounds that it is undeserved (e.g., caused by brute bad luck, see e.g. [22]), relational justice points out that these inequalities divide society and undermine the ability to interact based on mutual respect [23]. This theoretical account of relational justice builds on the foundational work of Iris M. Young [24] and Elisabeth Anderson [25], shifting the conceptualization of equality from material and distributive notions towards an inherently social and political value [26].

Thus, based on the topics that are consistently ascribed to the third wave of AI ethics, we hypothesize that the implications of the third wave for algorithmic fairness and justice are best characterized by topics typical of the philosophical discourse on *relational justice.* Particularly in the second wave, relational justice theories have been merely a side note in algorithmic fairness research [e.g., 27]. While contributions from the third wave address topics typical for relational justice, many contributions don't name it as such, missing the opportunity to situate their contribution within the broader philosophical discourse. Our hypothesis is further confirmed by recent works devoted to *relational AI ethics*, which have just been published during the publication process of this work [28, 29]. Against this background, we investigate the third wave of algorithmic fairness through the lens of relational justice theories by conducting a systematic literature review (SLR) and synthesizing insights from scientific contributions that, while potentially not explicitly framed within relational justice, implicitly engage with a relational lens. Specifically, we asked: *Which topics*, *including critical and constructive approaches*, *are discussed within the algorithmic fairness literature that pertain to relational algorithmic justice?*

Therefore, our contribution is twofold: First, we provide an overview of topics that are discussed within the algorithmic fairness literature that pertain to relational justice, characterizing the third wave of algorithmic fairness. Second, by highlighting the relevance of relational justice theories for algorithmic fairness research, we provide a theoretical overview to support interdisciplinary research and deduce important gaps in the current research from this enriched

[2] While these philosophical origins are often named by technical approaches to algorithmic fairness, it is criticized may contributions still lack an in-depth understanding of the moral foundations of philosophy and justice [17, 18].

theoretical foundation. Thus, strengthening relational justice helps both practitioners and researchers to expand the focus beyond algorithmic outcomes by structurally assessing crucial limitations of formal algorithmic fairness metrics. Furthermore, it might help to resolve some academic tensions that arise because different disciplines argue against a backdrop of different social justice theories without being explicit about this.

The remainder of the paper is structured as follows: Sect. 2 details the philosophical discourse that shifts from distributive justice theories to relational accounts of justice, which underlie our SLR. Section 3 provides details on our methodology for the SLR. Section 4 presents the results and topics identified, and Sect. 5 discusses limitations and recommendations for future research. Section 6 concludes.

2 Philosophical discourse on justice: distributive and relational accounts

In the following, we outline the shift in philosophical theories of justice from distributive to relational accounts[3] to illustrate a parallel development in algorithmic fairness research that developed from a formal, distributive second wave to a structural, relational third wave.

Both accounts share the premise that every person has an equal moral worth and, accordingly, equal basic rights, making equality central to justice, motivating the doctrine of egalitarianism [21]. Historically, egalitarianism was dominated by distributive frameworks, focusing on the appropriate distribution of benefits and burdens among members of society. A central concern is the question of which metric should ground distributive equality, often called the "Equality of What?" [36] debate. Accordingly, distributive justice may focus on different metrics by which inequality is assessed (in the philosophical jargon often referred to as 'currency' of equality), such as resources [37], opportunities for welfare [38], or capabilities [39]. Distributive theories commonly distinguish acceptable and unacceptable inequalities based on individual responsibility for them. Under the assumption that being worse off without one's fault is unjust [40], society is obligated to compensate for involuntary disadvantages [41]. For example, a child born with disabilities is not to blame for needing more help compared to healthy children; thus, society has the moral responsibility to resolve the resulting disadvantages. Institutions must ensure accessible buildings, for example, by providing ramps and elevators. Yet, those who caused their bad situation by themselves must accept their responsibility and bear the consequences of their decisions [22]. A self-inflicted bad does not justify entitlement to compensation or redistribution [25].

Relational egalitarians criticize this focus on individual responsibility and distributions. In 'What is the Point of Equality', Anderson [25] argues that distributive frameworks unjustly deny compensation and support to individuals by attributing suffering to personal fault [25]. This is called the 'harshness objection' [42], indicating that bad choices are punished too severely. More fundamentally, Anderson [25] criticizes the focus on compensation for presupposing that some personal qualities are inferior. For example, framing disability as a deficit requiring compensation obscures the role of social and institutional arrangements that produce exclusion. Focusing on compensation indicates that those affected by disabilities are inferior and themselves part of the problem, while in fact, the societal handling of disability is the problem. Material compensation alone does not remedy relational injustices such as deciding for the disabled without considering their voices, denying their agency, and failing to recognize them as equal moral participant in decisions that affect their lives. Thus, Anderson [25] argues that compensations are based on pity, and the unfortunate are stigmatized and disrespected. Similarly, Young [24] argues that distributive accounts are too narrow as they locate injustice primarily in unequal outcomes rather than in institutional structures that enable domination and oppression.

In response, relational theories of justice reconceptualize equality as a social and political value rather than a distributive pattern [26]. They treat unequal distributions as symptoms of deeper relational injustices rooted in domination, oppression, and structural power asymmetries. Justice, on this view, requires treatment with reciprocity and mutual respect, including the idea that no one should perceive

[3] Our focus on distributive and relational theories reflects the dominant development in the literature on algorithmic fairness. However, as an anonymous reviewer aptly pointed out, there are other notions of fairness in philosophy, including procedural, informational, and interpersonal formulations [30, 31]. In particular, distributive notions have received growing attention in the field of algorithmic fairness [32–34]. While we do not address these notions here, we agree that engaging with additional philosophical scholars and distinguishing their work from relational and distributive theories within the algorithmic fairness literature would be valuable for future work. Furthermore, it should be noted that, despite the different focus of relational and distributive egalitarianism, their exact relation is a complex contemporary debate as relational approaches have distributive implications and the other way around. For example, relational egalitarians acknowledge that a minimally sufficient distributive equality is a prerequisite to respect for another as equals [e.g., 25]. However unequal distributions are not the subject of justice in itself. Therefore, "Distributive and relational accounts of equality and justice are distinguished by their different starting points: patterns of distributions of advantages of all kinds […] or the quality of relations and interactions" [35, p. 620]. Relatedly, we do not suggest that the focus on relational justice within algorithmic fairness would render formal and distributive approaches to algorithmic fairness irrelevant. An unbiased algorithm would be desirable in certain contexts, but it also does not represent a sufficient analysis towards algorithmic justice.

themselves as superior or inferior to others, and that no one should be at the mercy of prevailing power structures. Relational injustice arises both interpersonally, for example, in a partnership [43], and institutionally, particularly through social processes that systematically constrain the capabilities of some groups while benefiting others [24]. These injustices result from aggregated, non-blameworthy processes, making it difficult to attribute individual responsibility in the event of harm [44]. The structural injustices that arise as a result of those processes become visible in the form of racism, colonialism, sexism, and so on, and define the background conditions of individual opportunities by influencing the options available, which either enable or constrain their actions [45].

The preceding section illustrates that relational theories mark a fundamental shift in the conceptualization of equality, opening room for new questions of justice that extend beyond a distributive framework. A comparable shift can be observed in the development of algorithmic fairness research, particularly in the transition from the second to the emerging third wave, as the second wave of algorithmic fairness is predominantly grounded in distributive ideals of justice [19]. In this context, a 'fair' algorithm is roughly interpreted as one that is not skewed or biased in favor of one group of people [e.g., 46]. This led to the adoption of an outcome-oriented perspective, operationalizing fairness through mathematical metrics designed to detect and mitigate unequal distributions of predictive outcomes across social groups. While several contributions lack an in-depth understanding of the moral foundations of ethical philosophy [17, 18], it is observable that many approaches are roughly linked to John Rawls as a philosophical inspiration for algorithmic fairness research [e.g., 47–49]. Rawls' theory of justice, especially his Difference Principle, is centrally concerned with distributive justice as it offers specific principles governing how social and economic benefits and burdens should be distributed to improve the situation of the least advantaged in society [50]. Within the third wave, a richer engagement with more diverse philosophical theories and a broader range of topics, including social relations and power inequalities, is observable. Some contributions that can be considered as part of the third wave explicitly demand overcoming the narrow, distributive frame. In this vein, for example, Kasirzadeh [19, p. 350] argues that …

> the exclusive focus on distributive accounts of justice invidiously deals with a single, narrowly construed aspect of justice. John Rawls who is one of the most influential contemporary political theorists and philosophers of distributive justice himself acknowledges that distributive justice addresses only a single instance of the application of the concept of justice [45, p.8]. Accordingly, the narrow scope of distributive justice carries over to algorithmic fairness.

Therefore, building on the philosophical background of distributive and relational theories of justice, we observe an increasing focus in the third wave of AI ethics on structural and relational injustices. Gaining a clearer picture of what might be overlooked within a purely distributive frame, the specific relational topics addressed in the third wave of algorithmic fairness are systematically synthesized by means of the following SLR.

3 Methodology

Our SLR is based on the well-established guidelines by Kitchenham [51] to ensure unbiased collection and synthesis of previous research [51, 52].

3.1 Preliminary sample

To address the research question, the search query had to capture the research area of algorithmic fairness combined with identifiers of relational justice. Formalizing the literature on algorithmic fairness, technical keywords "algorithm", "artificial intelligence", and "machine learning" are combined with the broad topic "in/justice" or "un/fairness". To identify the keywords for relational justice comprehensively, we used Nath's [53] *Philosophy Compass* survey on relational egalitarianism as an initial guide, supplemented by central literature on relational justice [25, 43, 44, 53]. As discussed above, relational theories challenge social hierarchies, including unequal status and power dynamics. To approach papers concerned with unequal social status hierarchies, we included the keywords "social hierarchies", "social status", and "social class". Because the term 'power' often refers to electrical power or performance, we specified the keywords "human power", "power dynamics", and "power hierarchies". Experience of privilege and oppression arises from the interaction of multiple social categories that shape lived experiences, positioning people differently within a world shaped by dominant systems of power [54]. As this is the main concern of intersectionality scholars, we decided to include "intersectionality" as a keyword. Unequal hierarchies may undermine relational equality because they enable domination [53] and oppression, which are at odds with respecting each other as equals. Therefore, we added "oppression" and "domination" to the search query. The disapproval of oppression also refers to oppressive states; thus, relational egalitarianism rejects undemocratic governments (ibid.). From this, a high value of democratic processes can be derived, and we decided to

include democratization and the related concept of participation. As mentioned above, the institutional understanding of relational justice pays particular attention to structural injustice, which has been included in the search query. Nath [53] furthermore highlights that relational equality is also expressed by feminist and critical theories, so we included the terms "feminism", "feminist", and "critical theories". Finally, we included broader terms that indicate the reference to relational egalitarianism (i.e., "relational ethics", "relational egalitarianism", and "relational equality") in the search query. Summarizing, we derived the following keywords for relational justice: structural (in)justice, feminis*, intersectional*, critical theories, social hierarchies, social status, social class, power dynamics, human power, power hierarchies, domination, oppression, participatory, democratization, relational ethics, relational egalitarianism, relational equality, social hierarchies, social status, or social class. The Boolean operator 'AND' connects the technical keyword, broad topic identifier, and a relational justice keyword[4]. If the predefined combination of search strings appeared in the title, keywords, or abstract of a publication, we included that publication in the preliminary sample for the following screening process.

To ensure that the SLR captures the variety of the interdisciplinary research field, we used multidisciplinary search databases (Web of Science, Scopus) and discipline-specific databases (ACM Digital Library, PhilPapers). We did not limit the timespan for the results. The predefined search query was applied to the four databases on the *01. August 2024*. Except for manuscripts not written in English or not being academic publications, all *676* results were included in the preliminary sample, which was the basis for the following screening process.

3.2 Screening

After removing duplicates, 400 distinct publications were left for the manual study selection. To this end, two researchers performed three screening rounds based on title, abstract, or full text. In case of disagreement, we engaged in mutual deliberation until a consensus was reached. When in doubt, the publication was kept for the next screening round. In the initial screening, we removed 163 publications based on titles that clearly referred to a different research area, so

[4] Full search string, adopted to the specific database website syntax: (algorithm* OR "artificial intelligence" OR "machine learning") *AND* (*justice OR *fairness) *AND* ("structural injustice" OR "structural justice" OR feminis* OR "critical theories" OR "power dynamics" OR "Human Power" OR "power hierarchies" OR "social hierarchies" OR "social status" OR "social class" OR "relational equality" OR oppression OR "domination" OR intersectional* OR participatory OR democratization OR "relational ethics" OR "relational egalitarianism" OR "relational equality").

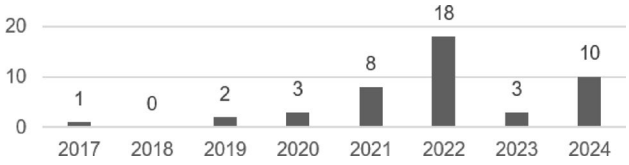


Fig. 1 Paper contributions per year, 2024 until August

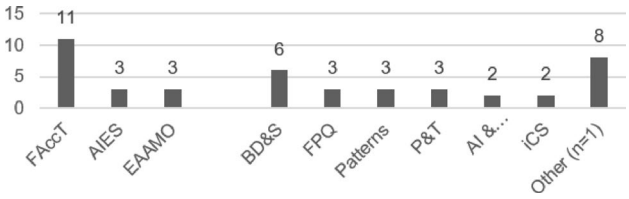


Fig. 2 Distribution of publications. left: conferences, right: journals

237 publications remained. Then, we formulated three guiding questions for the abstract and full-text screening: Does the paper engage with structural (in)justice, and/or does it refer to approaches from relational egalitarianism? Does the paper generally focus on justice or fairness evaluations, not only on a specific application? Does the paper discuss algorithmic systems/ machine learning/ AI while not referring to technology in general?

Based on these guiding questions, two researchers identified 72 publications that answer these questions positively.

In full-text screening, 28 publications were additionally excluded because they did not align with the guiding questions or our research question. Furthermore, the paper 'Data Feminism for AI' [55] was added manually due to its high relevance. It was not detected automatically, because it does not name 'Artificial Intelligence' but only the abbreviation 'AI'. While this might have excluded other relevant papers as well, we assume that abbreviations are most often explicitly named in the text, and 'AI' as a search keyword was not feasible due to its use in various contexts (e.g., AI=American Indians). This screening process resulted in a final selection of 45 publications [3, 14, 19, 20, 55–95], which are subject to in-depth analysis in Sect. 4.

3.3 Sample description

As illustrated in Fig. 1, with a deviation in 2023, topics within the scope of relational justice seem to be increasingly addressed. As the presented SLR was conducted in August, not all papers from 2024 could be included in the given sample. The chronological development of the identified publications highlights the increasing importance of the topic in recent years, with the first articles published in 2017. Figure 2 illustrates the distribution among conferences or journals. 18 papers are published at high-ranked conferences, influential for algorithmic fairness research. Furthermore, 27 papers were published in journals, with Big Data and

Society (BD&S) publishing the most. This was followed by Feminist Philosophy Quarterly (FPQ), Patterns and Philosophy & Technology (P&T), each of which published three papers. These conferences and journals are all characterised by an interdisciplinary orientation. This aligns with the interdisciplinary research topic, so the contributions cannot be clearly divided into distinct scholarly disciplines.

A high national homogeneity is observable, which might have been influenced by the restriction to English contributions. With 51,1%, most papers are written by authors working in the USA. Further, 40% of the documents originate in Europe, where the United Kingdom represents the most significant portion of papers (15,5%), followed by the Netherlands (8,8%), Germany (6,6%), and Ireland (6,6%). With two papers, Australia represents 4,4% of the sample, followed by one paper from Canada (2,2%) and one paper from Hong Kong (2,2%). All in all, it should be noted that this collection represents almost exclusively Western democracies, without any involvement of contributions based in South America or Africa, and only one originating from Asia.

3.4 Data analysis

The first two authors then investigated each paper in depth using a thematic analysis for qualitative data [96], drawing on Braun and Clarke's six-phase framework [97–99]. This includes several phases of analysis, such as familiarising oneself with the corpus, searching, reviewing, and, finally, defining and naming the themes. After the initial familiarization with the corpus (*phase 1*), we worked systematically through the data (*phase 2*). For our research efforts, this entailed the use of the software Obsidian, a Markdown-based knowledge management system that supports the creation of interconnected notes through internal links. Throughout summarizing the main findings of each paper, we manually created links whenever similar content was already addressed in another paper, in order to reveal relationships across the data and find initial codes. While summarizing the main findings of each paper, we placed additional emphasis on critical and constructive approaches, as defined by our research question. When searching for themes (*phase 3*), we did not predefine topic categories; instead, we allowed the topics to evolve from the corpus. This approach allowed us to identify, bottom-up, which components of relational justice are already addressed in the algorithmic fairness literature, rather than defining, top-down, which topics we are seeking. This phase was supported by Obsidian's graph view, which provides a visual representation of note interconnections. Although the process of linking the papers did not result in an organized thematic map, the resulting graph view was still useful for identifying and exploring thematic structures. After reviewing the themes (*phase 4*) by several rounds of critical and analytical discussion, we synthesized the overall findings, explained our reasoning to each other, and addressed any remaining ambiguities and discrepancies. To increase the internal homogeneity, we extended the six identified themes into eight themes and named them accordingly (*phase 5*). The identified eight recurring topics are synthesised and discussed in the following section, which represents the last phase (*6*) of Braun and Clarke's framework for thematic analysis.

4 Results

This section presents the results of the SLR in eight topics. Remarkably, some contributions explicitly acknowledge relational considerations [3, 59, 68], while several demand to overcome narrow and simplistic approaches to algorithmic fairness, by considering the societal context and background conditions of structural injustices [3, 56, 59, 61, 65, 66, 68, 94, 95]. For instance, authors distinguish "formal algorithmic fairness" [3, 95] or "fairness-as-parity conceptualizations" [19] from the philosophical accounts of justice, stressing that algorithmic fairness currently is restricted to distributive ideals [19, 57, 62, 72, 76, 100].

4.1 Power dynamics and epistemic injustices

Power structures that enable domination and oppression are central to any relational justice theory. In general, power can be described as the ability to make impactful decisions about the interests of others [86]. As the reasoning of *who should have* this ability is poorly addressed for algorithmic systems, it poses questions of legitimacy [86]. Instead, current practice focuses on the way ('*how'*) decisions ('*what'*) are made, leading to an exaggerated focus on principles such as accuracy or fairness [86]. The *'who'* is currently shaped by software companies controlling the algorithmic systems and determining the underlying standards and choices [92]. Significantly, this target group is represented mainly by a privileged group [59, 70, 74, 84, 89, 92]. AI research is predominantly Western-centered and based on cisnormativity and white supremacy, marginalizing the knowledge of those who are considered inferior [59, 67, 70, 74, 80, 81, 88, 89]. Thus, societal power structures are manifested in algorithmic systems [77], reflecting the interests of a homogenous and privileged group, and reinforcing injustices and existing social orders [65, 70, 74, 84, 92]. In turn, Miragoli [81, p. 1] conceptualizes AI "as a tool for epistemic oppression" because it functions as a gatekeeper of epistemic resources, mediating the provided contents.

The epistemic injustices are maintained due to the misconception that (natural and technical) science represents an absolute, objective, and value-free truth [20, 73, 88]. (Western) Science appreciates rationality above emotions [59, 73] and conceptualizes knowledge as stable, logical, and shaped by certainty [59]. Yet, the rational worldview is challenged by the feminist philosophy of knowledge, highlighting that scientists are subjective in their choices and assumptions, which affects their research [60, 73]. The dichotomy between reason and emotion reinforces dominant power hierarchies and devalues other forms of knowledge, such as lived experiences [55].

As a path forward, feminist theories of knowledge and epistemology are highlighted by some researchers [55, 59, 70, 74]. For example, Tacheva [88] advocates a *transnational* feminist approach. To do so, Gwagwa et al. [67] and Birhane [59] highlight Afro-feminism and the African philosophy of Ubuntu, and McFadden and Alvarez [79] stress the importance of critical self-reflexivity within AI ethics research. Feminist approaches focus on the marginalized experiences of those affected, as supported by many authors [20, 59, 61, 66, 70, 73, 74, 80].

Some authors identify new social classes based on epistemic injustices and power hierarchies. Cinnamon [64, p. 615] illustrates that society is divided into three 'data-classes' depending on their roles of creating, collecting, or analyzing data. This introduces a "data-class status hierarchy" which shapes the degree of control over individual lives. Similarly, Burrell and Fourcade [63] denote a 'coding elite' (p. 215), who strengthens its power through technical control and benefit from marginalized workers. On the adversary of the 'coding elite', Naudts [82, p. 5] notes a 'novel category of powerless' referring to "those without any authority or autonomy in the society's increased datafication" who are vulnerable to the privileged controlling user choices. This control can be misused for manipulation, for example, by triggering addictive behavior in online environments [63]. All this creates obstacles to participation in social structures, underrepresentation, and a digital ignorance of marginalized interests [82], turning algorithmic systems into a tool for domination over humans [62, 82]. These dynamics are further intensified by two distinct forms of invisibility that computer science practitioners operate within [79]: First, they benefit from historically and structurally produced systems of oppression, yet these advantages frequently remain unexamined and are implicitly encoded as normative assumptions within the technologies they design and deploy. Second, ethical considerations are rendered abstract and accountability is reassigned to imperfect (i.e., biased or discriminatory) machines, typically characterized as opaque or black-box systems, obscuring the responsibility of the practitioners in the first place (ibid.).

Summarizing, algorithmic systems affect, mediate, and change the social relationships between individuals and institutions [61, 73, 77]. They have a restructuring effect with a fundamental impact on individual (inter)actions and experiences of injustice [82]. Finally, Rafanelli [84] dismisses the fear that algorithmic systems might overpower humanity, arguing that AI is a tool for humans to exercise power, not a replacement for it.

To counteract these trends, McFadden and Alvarez [79] provide an overview of systems of oppression (capitalism and classism; colonialism and imperialism; racism and white supremacy; sexism and patriarchy; ableism), enabling AI researchers to build on the work of algorithmic oppression on a philosophical foundation. Tacheva and Ramasubramanian [87] conceptualize an analytical framework of an 'AI Empire' to highlight oppressive and interlocking power dynamics of AI that must be overcome.

Despite the criticism, it is acknowledged that algorithmic systems can also positively influence power structures. Park et al. [83, p. 17] observe an "often under-appreciated role of a power-balancer that algorithms can play". Then, algorithmic systems might be used to redistribute power, helping to supervise decision-making processes more closely (ibid). Klein and D'Ignazio [55] motivate the use of algorithmic outputs as an incentive for taking action towards social justice, rather than as a means to merely describe societal phenomena. Finally, Davis et al. [65, 1] sketch technology as potential "tools of liberation" that "can expose, undo, and reshape status quo".

4.2 Economic forces and capitalism

The development of algorithmic systems is driven by the logic of capital [78, 92]. However, authors highlight that the illusion of automation depends on crowd-workers' labor [63], which is usually poorly recognized and highly underpaid [63, 72]. Hierarchical and colonial labor intensifies poor work conditions, as workers risk losing their jobs if they don't follow the instructions [72], leading to involuntary acceptance of a toxic context and exploitation [82]. Underpaid labor contrasts with the high financial gains of companies, illustrating the domination of power asymmetries in the labor market [68]. Also, Burrell and Fourcade [63] highlight that extreme wealth and poverty coexist in regions where technology is produced. Further, due to colonial forces, exploitative data-labeling jobs are outsourced, and the Global South is more severely affected by the environmental impact of AI [55].

Additionally, capitalism and algorithmic justice intersect through companies' economic incentives in fairness research, potentially leading to conflicting interests of AI ethics experts due to their affinity with economic actors

[68], which also shapes the research agenda and determines which studies receive funding [62]. This prioritizes efficiency over maximizing justice in research on algorithmic fairness and may lead to conflicting strategies [73, 78, 88, 92].

Capitalistic dynamics also impact questions of data ownership, transferring personal data into a material good [64] that is sold for profit [57]. Those who produce the data and those who control, access, and gain money from it are separated, amplifying exploitative practices [64] and limiting access to data and computational infrastructure to a few companies [80]. With examples of the exploitation of homeless people and the use of images from child pornography or of dead people, Hampton [20] highlights questionable, disempowering, and nonconsensual data collection practices. These examples also underline the importance of autonomy over self-representation and privacy, protecting one's personality [56].

Relatedly, Braun and Hummel [61] demand the ability to control engagement and visibility in the digital data markets by oneself. Feminist critiques of consent demonstrate that the current approaches to consent rely on simplistic, binary, and overly individual-focused assumptions, which are not suitable for AI [55]. A proposal to include feminist principles has been suggested by Ajmani et al. [57], who developed a 'data agency theory' based on Young's conception of social justice. They stress the "systemic relationship between consent and agency in data settings" and conceptualize "data agency as a necessary dimension of justice in predictive systems" (p. 632), referring to the individual's capacity to control the use of the data they create.

4.3 Obstacles to social solidarity

Focusing on solidarity is discussed as a path toward more algorithmic justice by some authors. For instance, Braun and Hummel [61] consider solidarity as "shared practices of individuals or groups based on perceived similarities, in particular, in connection with their views and goals" (p.5) and reemphasize the commitment to social solidarity to eliminate algorithmic injustice. Applying solidarity to algorithmic systems implies that shared practices are developed, all voices are heard, and those marginalized are focused (ibid.). Klein and D'Ignazio [55] stress the importance of solidarity across different social groups to counteract capitalistic power and colonialism. Based on the philosophy of Ubuntu, Gwagwa et al. [67] emphasize that solidarity and relational approaches are vital for AI ethics. By incorporating the value of solidarity with a focus on communal relations, which are "based on generosity, hospitality, compassion, and friendliness" [67, p. 2], technology would be able to capture the value of communities and relationships more faithfully (ibid.).

However, although social solidarity can serve as a means to promote algorithmic justice, achieving this is difficult because algorithmic systems themselves can create obstacles to social solidarity across different contexts. For example, the described exploitative labor conditions isolate workers, hindering their collective solidarity, as they have no means for collective mobilization [63]. Algorithmic systems change how individuals are embedded in relationships with individuals and institutions [73]. Intensifying this, algorithmic predictions may disturb the comprehension of how personal success is hindered, leading to self-blaming instead of challenging the algorithmic systems, which makes injustices abstract and less actionable [73]. For example, if a hiring algorithm consistently rejects candidates from certain backgrounds, people may internalize the rejection as a personal failure rather than recognizing systemic bias. Given that the algorithm is assumed to be neutral and due to its opaqueness, it is hard to voice complaints or defend oneself, making the structural injustice harder to identify and address.

In yet another context, where algorithmic systems are used for content recommendation, individuals are treated as group representatives, and their engagement with (for their group) atypical content is reduced [73]. Due to less exposure to different groups, they become increasingly separated, and solidarity across groups becomes less likely. Cinnamon [64] argues that this separation hinders the free development of identity as online choices are narrowed through the best-fitting filter bubble. Furthermore, since algorithmic systems rely on correlation and rarely consider causal factors, people are evaluated based on statistical quantities instead of their actual actions or intentions [56]. As a result, algorithmic predictions cannot be empathic and fail to consider the uniqueness of individual experiences [73]. Finally, algorithmic systems cannot acknowledge uncertainty or structural injustices and may not give the benefit of the doubt [95], undermining social solidarity.

4.4 Techno-solutionism

Several authors argue that technical fixes to algorithmic systems' harms are not sufficient in tackling the root causes of social injustices [19, 68, 72, 73, 87]. Consequently, many technical overlook that social bias cannot be 'fixed': The idea of 'fixing' unfairness arises from the focus on debiasing algorithmic systems [59, 75, 76], reducing systems of oppression and structural injustices to the narrow dimension of algorithmic bias [20, 58]. In turn, minor mathematical changes toward debiasing might legitimize the use of an algorithm [19] while overlooking its actual societal impact.

Thus, algorithmic fairness approaches are susceptible to abuse, working under a "veneer of objectivity" [59, 95] and "render decisions seemingly objective" [65, 4], distracting attention from more urgent points of justice [72, 79]. For example, instead of finding the right categories, intersectional oppression must be addressed [72], and it is essential to center the purpose [75], especially since a technical fairness constraint is applied only to specific points of a process [90].

The premise of 'fixing' injustice indicates an orientation towards techno-solutionism, as criticized by several authors [3, 20, 58, 66, 68, 82]. This view limits the problems considered and thus also the available solution strategies [62]. Challenges without a clear solution strategy, structural reforms, and social inequalities are likely to be considered outside the scope [3, 60, 92]. Focusing on technical fairness makes it harder to challenge decisions on whether automated systems should be implemented [62]. Several authors question the beneficence of algorithmic systems in the first place, enabling non-algorithmic solutions and the option to disregard algorithms [3, 20, 62, 66, 68, 76, 92]. Zimmermann and Lee-Stronach [95] remember to consider the full range of available options: An algorithm usually does not indicate uncertainty when facing a boundary decision, as it is designed to give a binary prediction (yes or no), but there might be other actions that should be considered.

To overcome techno-solutionism, Green [3] demands a methodical reform, shifting from abstract, technocentric formalizations towards substantive evaluations. Therefore, he suggests a relational response that rejects social relations as neutral but instead addresses unequal opportunities and privileges that exist before the algorithmic decision and feed into it (referred to as upstream influences). He argues that this needs to be accompanied by a structural response that focuses on the downstream effects, such as the individual consequences of being judged unfavorably by an algorithmic decision, to reduce its potential harmful impacts after a decision was made (ibid.).

4.5 Categorizations

In current algorithmic fairness practice, the social construction of categories like gender or race is often ignored, treating them as objective facts [3, 74]. Social values are reduced to calculable categories [58, 73, 95], leading to flawed conclusions. For example, Rafanelli [84] warns that hiring software could falsely equate 'qualified' with 'like successful men.' By relying on these epistemic misassumptions, an algorithmic system "naturalizes group differences in risk that are the product of oppression" [3, p. 12] and "perpetuates a current system of differentiation" [73, p. 18]. Consequently, algorithmic systems prescribe societal norms and obscure transformation [62]. Furthermore, Lu et al. [78] argue that the fixed, non-relational categories immobilize conceptions of human identity. Structures of injustice become normalized due to data-driven decisions [82]. In turn, algorithmic systems entrench social norms by relying on socially constructed metrics that reproduce social hierarchies, thereby forcing individuals to adapt their behavior to the algorithmic evaluation [73]. Importantly, these do not follow any reasoning but are built on the tendency to "endorse attitudes and behaviours that are commonly accepted simply because they are commonly accepted" [81, p. 6]. Miragoli [81] describes this as 'epistemic conformism', mimicking the training data's emerging patterns.

A further challenge is that most approaches to algorithmic fairness focus on a predefined list of protected categories (e.g., gender, race). However, defining the categories that are needed in a certain context is not straightforward, given the vast number of potential options [72]. Fairness metrics are limited to legally protected and statistically observable categories, whereas certain forms of oppression remain excluded (ibid.). Thus, Kong [72] concludes that defining relevant protected categories is a political matter rather than a purely statistical problem. Questioning the legal scope, Aizenberg and van den Hoven [56] note that algorithmic systems enable wrongful discrimination that is not directly related to known protected grounds. Furthermore, unobserved characteristics unavailable within the dataset raise a particular challenge for protected categories [89]. An example is the social category 'class', which is rarely considered in algorithmic fairness research, while persisting in privileges across generations [89, 92].

Without being observed, these categories can hardly be protected. However, protecting privacy is equally important and poses a challenge to the current approaches to algorithmic fairness [58]. Privacy is crucial for vulnerable groups, as information disclosure can put them at risk, leading them to prefer invisibility in the data to avoid punishment [82]. An example is the disclosure of homosexuality, criminalized by several countries [89]. Naudts [82] concludes that unlimited data collection may not solve the problem of fair representation. Instead, "the interests of those who wish (or have to) remain silent and invisible" [82, p. 5] must be considered. Furthermore, data collection for algorithmic fairness might be misused to expand the surveillance infrastructure [56]. Surveillance structures lead to harmful hypervisibility of those oppressed, while they are made invisible simultaneously due to algorithmic biases and missing representations [20].

Protected categories are also criticized due to their stigmatizing effect, associating individuals and groups with inferiority and reinforcing "unjust processes of differentiation" [73, p. 20]. The incomplete picture of individuals

hinders the recognition of their unique talents and characteristics, harming the protection of human diversity [73]. This oversimplification misconceptualizes who represents which group and counts as a similar case [58, 95], reducing autonomy over personal identification. Algorithmic systems treat individuals according to (assumed) best-fitting category and may reinforce "oppressive or overly prescriptive categories" [58, p. 1715], risking to create isolated categories without considering their substantive meaning (ibid.). Categorizing humans or groups can create boundaries that limit people's actions and possibilities [82]. Naudts [82] emphasizes that when people are at risk of exclusion based on definitions imposed on them, the practice of classification itself should be questioned.

It is argued that categorizing humans is inherently flawed because fluid subjects, such as gender or race identity, which depend on the social context and may change over time, cannot be squeezed into discrete categories to be computable [58, 74, 89]. For example, the use of discrete categories fails to account for non-binary identities and trans people [89]. Everyone outside the categorization becomes invisible or summarized as 'other' [82]. Binary classification systems preserve oppression by obscuring hierarchies, and hence it is argued that "gender binary has been operationalized and even weaponized in AI research" [55, p. 104].

Human identity is far more complex and dynamic than is often acknowledged in AI research [78]. Identity is considered a mutually exclusive and fixed category across contexts, which assumes an inherent ground truth while neglecting the fact that this may not represent the complex nature of human identity [78]. Lu et al. [78] stress that understanding identity as a system of relational processes is crucial, and develop two technical approaches based on multilevel optimization and relational learning.

4.6 Intersectionality

It is argued that algorithmic intersectional fairness focuses on categories such as race and gender, but may not appropriately address the intersection of oppression [72]. However, algorithmic intersectionality must be analyzed in its relationship to power [14, 72, 77], considering the roots of colonialism and racial capitalism [87]. The dominant orientation toward statistical parity among subgroups [72] symmetrically weights the disadvantages between these subgroups while ignoring that intersecting disadvantages mutually intensify each other and may not be summed up [14, 73, 77, 95].

In their analytical paper, Davis et al. [65] argue that all three common fair ML approaches (anti-classification, classification parity, and calibration) are a mismatch to theories of intersectionality. Further, Kong [72] illustrates a conceptual dilemma when working with subgroups: either the developers have to split the population into an endless number of subgroups, or stop this split somewhere, leading to the arbitrary selection of protected groups [72]. A power-based analysis of intersectional algorithmic fairness reveals the exclusive focus on disadvantage but not privilege [14]. This is concerning because "by centering disadvantage, we fail to question the normative conditions that produce—and promote the qualities or interests of—advantaged subjects" [14, p. 907].

To counteract this, Lopez [77] proposes the susceptibility to algorithmic disadvantage framework based on a vertical dimension (relation state actor – individual) and a horizontal dimension (relation individual – other individuals). The vertical dimension highlights the vulnerability and dependency of (marginalized) individuals, whereas the horizontal dimension focuses on the social structure between those intersectional disadvantaged (ibid.). Another constructive approach is provided by Davis et al. [65], proposing the focus on 'algorithmic reparation' to address intersectionality in machine learning.

4.7 Affirmative action and forward-looking responsibility

Given the disparate impact of algorithmic systems across individuals and the structural nature of discrimination, several authors argue that the focus on statistical parity between subgroups is overly narrow [3, 58, 71, 72, 76, 94, 95]. In turn, algorithmic justice should focus on disparate impact rather than treatment, and several authors conclude that fairness metrics and affirmative algorithms in favor of marginalized groups may be justifiable, thereby providing a rationale for prioritizing certain groups. Davis et al. [65] propose a reparative approach grounded in affirmative weights. Kong [72] argues for interventions that increase the likelihood of positive outcomes for marginalized groups to reduce the effects of structural oppression. Similarly, Kassam and Marino [71] motivate different risk thresholds based on collective responsibility. Zhang [94] focuses on basic capabilities that must be secured to relate to each other as equals as humans, as participants in a system of social cooperation, and as citizens. As algorithmic decision-making systems alter the access to these basic capabilities, this may be a rationale for using different risk thresholds and affirmative algorithms (ibid.). Even if some might argue that unequal treatment is procedurally unjust, Zhang [94] counters this critique by pointing out that the orientation for substantive justice might justify different treatment. While not directly referring to affirmative algorithms, Braun and Hummel [61] also agree that a "prioritization of those who have remained insufficiently considered up to now" (p. 6) is needed. While

acknowledging the rationale for affirmative action, the most critical point is highlighted by Green [3]. He emphasizes the dilemma that treating everyone equally may reproduce inequality, whereas special treatment may stigmatize the respective groups. In this way, affirmative treatment might also fail to enable greater equality.

An alternative approach to addressing structural algorithmic injustices focuses on responsibility rather than fairness. Addressing the question of responsibility, Burrell [62] stresses the importance of avoiding the shifting of responsibility to consumers. As an alternative approach, Lin and Chen [76] highlight the social connection model – an approach of forward-looking responsibility by Iris Young [45] and argue for collective action, addressing responsibility beyond software engineers and acknowledging that different actors have different capabilities and roles to reduce structural injustices. From a similar perspective, Zhang and Hosoi [93] illustrate parallels between algorithmic systems and Young's account of structural injustice by stressing that they "can reify status quo inequalities without being biased or discriminatory in a traditional sense, and without being the result of blameworthy behavior by a discrete bad actor" [93, p. 2].

The misleading narrative of 'bad actors' is produced by the predominant focus on faulty individual perpetrators and their intentions in many anti-discrimination laws, ignoring the broader structural and social context that caused the discrimination in the first place [14]. This carries over to algorithmic systems by focusing on algorithmic biases, technical causality, or specific blind spots in the design process rather than foregrounding "the ways humans and technology co-conspire to not just passively reproduce but actively uphold and reproduce discriminatory social structures" [14, p. 904]. Focusing on bad intentions is also misleading because, as Iris Young [45] argues, structural injustices arise from the aggregation of non-blameworthy actions. In this line, Kong [72] summarizes: "Most engineers do not deliberately discriminate against Black women. The problem is rather that they are simply doing their jobs, but their actions contribute to reproducing oppression" (p. 488). This example shows that focusing on intention is problematic when facing algorithmic discrimination. Correlations in the data may lead to proxy discrimination, which happens without intention or even without being noticed [71]. These points are further continued by Kasirzadeh [19], highlighting some practical consequences of the social connection model for algorithmic justice research. Her analysis reveals the need for careful evaluations of the goals when developing algorithmic systems for a social context, the focus on the interplay between algorithmic predictions and social power, and the attention to non-technological trade-offs with increasing interdisciplinary analyses. Another practical approach is proposed in Naudts [82], demanding the promotion of 'cultures of reflexivity' to assess responsibility across actors and institutions instead of blaming individuals or single departments.

4.8 Democratic and participatory approaches

Democratic and participatory approaches to algorithmic systems are a trending topic [60, 69]. However, contradictory claims are evident, as summarized below.

Both instrumental and instinct values are allocated to algorithmic democratization and participation. On the instrumental side, participation shall help to include various forms of knowledge [59, 70, 74]. Community-based auditing methods can uncover biases that would have been missed otherwise [90]. Enabling direct and self-determined control over group identities can address misrepresentations [58]. Kong [72] suggests collaborative projects between communities and researchers to foster shared responsibility and to test high-stakes algorithms with diverse stakeholders.

On the intrinsic side, democratization and participation are discussed as a matter of justice. The lack of participation results in opaque decisions [82]. Cinnamon [64] illustrates that dominant data practices exacerbate three obstacles to parity in participation (maldistribution, misrecognition, and misrepresentation). Meaningful participation requires including the voices of those affected [68], letting the decision-subjects' needs guide the development of algorithmic systems [92], enabling self-determination and community empowerment [59] and building on various perspectives by engaging with impacted communities [55].

To enable algorithmic participation, several methods are introduced. Aizenberg and van den Hoven [56] offer an overview of Design for Values, Values-sensitive design, and Participatory Design. Furthermore, the concept of 'design justice' by Costanza-Chock [101] is recommended by several authors [20, 59, 68, 74, 92]. To realize the democratization of AI, Wong [91] suggests an accountability for reasonableness framework that aims to be "reason-giving and reason-responding […] that show respect to individuals' views and voices and recognize their differences" [91, p. 239]. This requires the use of precise language, communicating approaches and weaknesses of algorithmic fairness, focusing on marginalized perspectives, and developing strategies to address conflicting views on algorithmic systems [91]. Medrado and Verdegem [80] focus on Latin American participatory action research to challenge "how the North's centrality is taken for granted regarding AI epistemologies, experiences, and understandings" (p. 1). This involves incorporating bottom-up approaches and actively supporting the colonized or oppressed (ibid.). Naudts [82] proposed recommendations to enhance engagement in the politics of the digital society, including recognizing diverse

social roles, examining power and privilege in AI development, and embracing multiple perspectives.

Critical contributions on participation and democratization of algorithmic systems question its implementation [69, 85]. Himmelreich [69] argues that "more democracy does not mean more participation" (p. 1337). As the ideals of democratic or participatory algorithmic systems are abstract, lacking defined goals or methods, calls to democratize algorithmic systems are without a clear substance and missing a distinction between participation and deliberation [60, 69]. Instead, standards for participatory algorithmic systems should include the level of reciprocity, reflexivity, empowerment, and task duration [60]. Furthermore, the democratic boundary problem is highlighted, questioning how to include *all* members of society in the design and use of algorithmic systems [69]. Birhane et al. [60] question the industry's interest in participatory algorithmic systems and emphasize the importance of non-experts.

Finally, the superficial conception of algorithmic participation and democratization makes them susceptible to misuse. Sloane et al. [85] coined the term 'participation-washing', highlighting the risk of exploitative community involvement. Companies may benefit economically from misusing participation as work [85]. Similarly, Himmelreich [69] cautions against the ideological exploitation of participation, as it might be misused to hide the centralization of power. Participatory approaches often ignore and thus fail to fight power asymmetries [59, 60, 69, 85]. Furthermore, participatory research is primarily driven by privileged perspectives [75]. Thus, a grounded relationship between researchers and communities is essential, shifting the role towards "a trusted partner and advocate" [75, p. 564]. Finally, to use participatory research results constructively, engagement in value transparency is needed, i.e., analyzing the rationale behind the decisions and choices [83].

5 Discussion

Through the conducted SLR, we obtained eight topics of current research and gaps in the algorithmic fairness literature that pertain to relational justice. Overall, most papers critically engage with the limitations of dominant formal fairness approaches, emphasizing that algorithmic justice cannot be achieved by technical means alone. This confirms Branford and Herzog [29] in their argumentation, relational AI ethics offers a nuanced lens for identifying and defining AI-related ethical issues, but its potential is less realized in driving and operationalizing concrete solutions. Yet, it should also be noted that several papers already offer promising **constructive approaches**, such as the 'data agency theory' based on Young's conception of social justice [57], the suggestion to focus on 'algorithmic reparation' to address intersectionality in machine learning [65], or focusing on the Social Connection Model of collective responsibility by Young [45] as emphasized by Lin and Chen [76] and Kasirzadeh [19]. Beyond this, this SLR identified several promising directions for **future research**, as discussed in turn.

First, the section on *power and epistemic injustice* highlights unresolved questions about who holds legitimate authority in decision-making processes involving algorithmic systems. Insufficient attention has been paid to questioning the authority underpinning algorithmic fairness research. Given the impact of big-tech companies and their interest in the algorithmic fairness research agenda as discussed above, addressing rightful legitimacy is becoming even more crucial. Rightful authority (i.e., legitimacy) is, among others, closely tied to the fulfillment of ethical and functional obligations (i.e., responsibility). Both concepts focus on a specific 'who' question: for one, 'who should bear the responsibility?', for the other, 'who has the rightful authority for certain decisions?'. If these questions remain unaddressed for the algorithmic decision procedure or its outcomes, a 'legitimacy gap' [102] or, respectively, a responsibility gap [e.g., 103–105] might arise for algorithmic systems. Both questions are highly relevant for algorithmic fairness research that aims to challenge power imbalances and support the relational demands of mutual respect. Therefore, we suggest that the engagement with legitimacy and responsibility gaps in AI can be productive for a relational account of algorithmic fairness research. Especially, responsibility gaps in AI offer well-developed accounts to analyze and handle different kinds of these gaps [103, 104] and might, for example, be beneficial for transferring the implications of the social connection model of forward-looking responsibility for justice discussed above to the specific context of algorithmic systems.

A further key takeaway from the highlighted epistemic injustices is the need for future research to include other forms of knowledge besides rational ones, such as embodied [106] or tacit [107] knowledge. Broadening the conceptualization of what constitutes relevant knowledge aligns with the aims of relational justice, which emphasizes the importance of respectful relationships, mutual respect, and recognition. Limiting knowledge to rational or expert-driven epistemic forms would undermine the relational aim, as it reproduces existing hierarchies by excluding the lived experiences of marginalized groups. Future research that recognizes the various forms of knowledge would greatly benefit the discourse on algorithmic fairness. This might also include analyzing epistemic injustice as part of algorithmic audits to make this topic more strongly embedded in the AI pipeline, rather than a critical afterthought. The relevance

of epistemic injustices for algorithmic fairness research is related to two other core areas identified in the SLR. First, a purely rational perspective often overlooks the importance of solidarity, as discussed in Sect. 4.3. Engagement in solidarity offers both epistemic and relational advantages [108]. Relationships that build on solidarity regard one another as trustworthy, which enables shared information, joint deliberation, and mutual learning [108]. While (technical) research on the social impact of algorithmic systems tends to be dominated by values such as fairness, transparency, and accountability, incorporating solidarity more strongly into the discourse could become a means of promoting justice and recognizing marginalized perspectives (see Sect. 4.3 [55, 61, 67]), . Future research is needed to understand how this can inform current practices and create additional opportunities for building solidarity in AI.

Second, to reduce epistemic injustices, *democratic and participatory* (Sect. 4.8) approaches might enhance the algorithmic fairness research. The SLR results showed that increased participation could lead to shared responsibility, identification of stakeholder needs, self-determination, and knowledge diversification. Yet, the assessment of the potential of democratic or participatory approaches to algorithmic fairness is inconclusive. The need to define the ideals, goals, and methods of democratic or participatory AI is critically emphasized and represents a critical gap in the literature. Extended conceptual research would protect the aims of participatory and democratic proposals against misuse and participation washing. Furthermore, as highlighted by relational justice theories, participation alone is not a sufficient substitute for citizen empowerment [24]. Therefore, future research, informed by a stronger philosophical grounding, is needed to align participatory AI projects with the relational ideals that underpin them. In this regard, we recommend a contribution published shortly after we conducted this SLR. Maas and Moreno Inglés [109] argue for critical engagement beyond participatory AI and assert that, for effective empowerment, participation must reduce the arbitrariness of power relations between a system's developers and its stakeholders. Finally, future research on democratic AI must critically examine which form of democratic deliberation it wishes to support (see Santoni de Sio [110], Chaps. 10–12). In doing so, it takes into account the findings on epistemic injustices discussed above.

The success of these approaches is also dependent on their institutional integration. As highlighted in Sect. 4.2, the development of AI and research on algorithmic fairness are strongly influenced by economic forces and the capitalist market. Given their influence on algorithmic fairness research, attention to economic power is especially relevant and is also intertwined with the legitimacy gaps discussed above. To determine how future research on algorithmic fairness should be governed, it is necessary to disentangle the economic power dynamics that shape this research. Who should influence the algorithmic fairness research agenda? What and whose interests and goals does algorithmic fairness research address? How can this be better aligned to foster social outcomes rather than merely increasing efficiency and legal compliance?

Addressing these questions may help avoid uncritically reinforcing techno-solutionism while remaining receptive to the potential benefits of algorithmic fairness strategies. It is crucial to navigate between the contributions of algorithmic fairness research while remaining critical about its societal limitations. The initial step must always entail a thorough examination of the appropriateness of using a specific algorithm in a given context. The adherence to specific technical metrics does not constitute proof of fairness. Instead, algorithmic fairness metrics are a powerful tool for proving *un*fairness by demonstrating the biased characteristics of specific applications. As such, research on algorithmic (un)fairness provides strong data for legal cases and governmental decisions. It is not only the limitations of technical fairness formalizations, but rather their governmental use that should be subject to greater discussion in future research. A suggested answer is to adopt affirmative action (Sect. 4.7), i.e., to actively use bias-mitigation metrics to combat injustice. Building on this, future research at the governmental level is needed to understand how the technical reduction of bias may alter the status quo.

Although not all reviewed papers explicitly engage with relational theories, many implicitly address their core concerns, demonstrating a growing interest in relational ideas within the field. Given the theoretical complexity of relational justice, each topic presented is only one component on the path toward relational justice. Aiming for relational algorithmic justice should not be reduced to simplistic formalizations, and sustaining its complexity is essential. This theoretical and conceptual complexity also imposes **limitations** on the present SLR. Given the broad scope of relational and adjacent philosophical discourses, the review focused on a targeted set of keywords that signaled engagement with relational themes. This necessarily narrowed the range of included studies. While the selected papers align thematically with relational justice, they do not encompass the entire discourse. Instead, the SLR should be understood as an initial step toward mapping the relational turn in algorithmic justice. It is important to acknowledge that the deliberate focus on the research area of algorithmic fairness and justice narrowed the scope of this SLR. Thus, it is essential to emphasize that other topics, such as research on climate AI and its manifested neocolonial and racist power dynamics [111] and forces of colonialism [112, 113], are also highly relevant for a relational justice analysis in the

context of AI. Thus, the interplay among climate justice, colonialism, relational justice, and algorithmic systems warrants further investigation.

In general, our synthesis of the findings revealed a dominant focus on the work of political philosopher Iris M. Young [19, 57], [e.g., 61, 62, 72, 76, 82, 84]. Despite her influential standing, future research should aim for deeper engagement, considering a broader range of philosophical schools. For example, relational injustices manifest not only at the institutional level [25, 44] but also in interpersonal relations [43]. In critical self-reflection, we acknowledge that, for this SLR, we focused primarily on institutional dimensions and a limited set of relational authors. Given that power hierarchies are one of the central concerns of relational injustice, other sources such as French and Raven [114][5] on the forms of power, Pfeffer and Salancik [115][5] on resource dependency theory, Foucault [116] on disciplinary power, or Hannah Arendt's [117] theory of power and communication. While further broadening the theoretical conceptualization was beyond the scope of this paper, we see this as a valuable opportunity for future research to incorporate additional philosophical concepts into the discourse on relational algorithmic fairness. The transnational feminist approach advocated by Tacheva [88] is a salient example in the results of this SLR, underscoring the need to avoid overreliance on a particular Western philosophical scholar. Furthermore, Gwagwa et al. [67] and Birhane [59] highlight Afro-feminism and the African philosophy of Ubuntu.

Additionally, the findings within this SLR are not closely analyzed with respect to the various technical fairness metrics proposed. Some of the identified takeaways apply to all proposed metrics, such as the need to enhance democratization to foster justice. However, other points, such as the limitation of intersectionality due to the focus on subgroups, are particularly pertinent to group-based approaches [118]. While individual fairness notions [47] might partly avoid this shortcoming[6], they face the challenge of defining what constitutes similarity, which complicates the categorization and epistemic injustice issues discussed above. Future research should analyze the different fairness metrics and their specific implications from a relational justice lens.

6 Conclusion

In the philosophical discourse, relational theories emphasize unequal power asymmetries, social relations, and structural injustices. Similarly, the third wave of algorithmic justice is described as a shift from distributive and formal metrics towards conceptualizing algorithms as sociotechnical systems and their impact on power asymmetries and structural injustices. To investigate what the relational perspective adds to the—so far mostly distributive—discourse on algorithmic justice, the presented systematic literature review synthesized insights from 45 papers into eight subject areas. Relational justice offers a theoretical foundation for future interdisciplinary work. By foregrounding social relationships, power dynamics, and structural inequalities, relational approaches overcome the pitfalls of techno-solutionism and instead promote a contextual and situated understanding of structural injustices. These perspectives enrich ongoing discussions of epistemic injustice, intersectionality, and accountability within algorithmic systems and suggest promising directions for future work. Thus, this systematic literature review demonstrates the growing importance of relational justice in the context of algorithmic fairness and suggests that a nuanced investigation of relational theories offers fruitful inspiration for future research.

Author contributions The research and writing were led by L.W., who conceptualized the research, developed the central argument, drafted the manuscript, and coordinated the project. M.M. contributed through in-depth discussions on structure and content, helping to refine ideas and provide critical feedback. Both L.W. and M.M. equally conducted the double review process. C.L. supervised the project, providing guidance and oversight throughout. All authors actively participated in reviewing, editing, and finalizing the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Chouldechova, A.: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data. **5**(2), 153–163 (2017). https://doi.org/10.1089/big.2016.0047

[5] We thank the anonymous reviewer for this accurate suggestion.

[6] Thanks to the anonymous reviewer for pointing this out.

2. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent Trade-Offs in the Fair Determination of Risk Scores. (2017). https://doi.org/10.48550/arXiv.1609.05807
3. Green, B.: Escaping the impossibility of fairness: from formal to substantive algorithmic fairness. Philos. Technol. **35**(4) (2022). https://doi.org/10.1007/s13347-022-00584-6
4. Crawford, K.: Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven, CT, USA (2021). https://doi.org/10.12987/9780300252392
5. Bolte, L., van Wynsberghe, A.: Sustainable AI and the third wave of AI ethics: a structural turn. AI Ethics. (2024). https://doi.org/10.1007/s43681-024-00522-6
6. Häußermann, J.J., Lütge, C.: Community-in-the-loop: towards pluralistic value creation in AI, or—why AI needs business ethics. AI Ethics. **2**, 341–362 (2022). https://doi.org/10.1007/s43681-021-00047-2
7. Kind, C.: The term 'ethical AI' is finally starting to mean something VentureBeat. (2020). https://venturebeat.com/ai/the-term-ethical-ai-is-finally-starting-to-mean-something/. Accessed 7 Jun 2023
8. Pessach, D., Shmueli, E.: A review on fairness in machine learning. ACM Comput. Surveys. **55**(3), 1–44 (2022). https://doi.org/10.1145/3494672
9. Caton, S., Haas, C., Survey: ACM Comput. Surv. **56**(7), 1–38 (2024). https://doi.org/10.1145/3616865
10. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–42 (2019). https://doi.org/10.1145/3236009
11. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., Explainable, A.I.: A Review of Machine Learning Interpretability Methods. Entropy (Basel Switzerland). **23**(1) (2021). https://doi.org/10.3390/e23010018
12. High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI, Brussels (2019)
13. Mitchell, S., Potash, E., Barocas, S., D'Amour, A., Lum, K.: Algorithmic fairness: choices, assumptions, and definitions. Annu. Rev. Stat. Appl. **8**(1), 141–163 (2021). https://doi.org/10.1146/annurev-statistics-042720-125902
14. Hoffmann, A.L.: Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Inform. Communication Soc. **22**(7), 900–915 (2019). https://doi.org/10.1080/1369118X.2019.1573912
15. Hamilton, I.A.: Why it's totally unsurprising that Amazon's recruitment AI was biased against women. Business Insider. (2018). https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10?r=DE&IR=T. Accessed 21 May 2021
16. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. In: Ethics of data and analytics, pp. 254–264. Auerbach (2022)
17. Binns, R.: What Can Political Philosophy Teach Us about Algorithmic Fairness? IEEE Secur. Priv. **16**(3), 73–80 (2018). https://doi.org/10.1109/MSP.2018.2701147
18. Lee, M.S.A., Floridi, L., Singh, J.: Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. AI Ethics. **1**(4), 529–544 (2021). https://doi.org/10.1007/s43681-021-00067-y
19. Kasirzadeh, A.: Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. In: AIES '22: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery (ACM), pp. 348–356 (2022). https://doi.org/10.1145/3514094.3534188
20. Hampton, L.M.: Black Feminist Musings on Algorithmic Oppression. In: Conference on Fairness, Accountability, and Transparency (FAccT '21), Virtual Event, Canada (2021). https://doi.org/10.1145/3442188.3445929
21. Arneson, R.: Egalitarianism. In: The Stanford Encyclopedia of Philosophy, Edward N. Zalta, (eds.) 2013th ed.: Metaphysics Research Lab, Stanford University [Online]. (2013). Available: https://plato.stanford.edu/entries/egalitarianism/
22. Gosepath, S.: Equality, The Stanford Encyclopedia of Philosophy Edward N. Zalta (ed.), (2021). https://plato.stanford.edu/archives/sum2021/entries/equality/
23. Miller, D.: Justice. In: The Stanford Encyclopedia of Philosophy, E. N. Zalta, (ed.) (2021). Available: https://plato.stanford.edu/entries/justice/. Accessed 13 Jun 2023
24. Young, I.M. (ed.): Justice and the politics of difference. Princeton University Press, Princeton (1990)
25. Anderson, E.: What is the Point of Equality? Ethics. **109**(2), 287–337 (1999). https://doi.org/10.1086/233897
26. Voigt, K., Wester, G., Relational Equality and Health: Soy. **31**(2), 204–229 (2015). https://doi.org/10.1017/S0265052514000326
27. Binns, R.: Fairness in Machine Learning: Lessons from Political Philosophy. In: Proceedings of Machine Learning Research, New York, pp. 1–11 (2017). [Online] Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3086546
28. Dignum, V., Building Inclusive, A.I.: Integrating Relational Ethics with Compositional Design, INESC TEC Science & Society. **1**(7) (2025). [Online] Available: https://science-society.inesctec.pt/index.php/inesctecesociedade/article/view/139
29. Branford, J., Herzog, C., Relational, A.I., Ethics: Strengths, Challenges, and Future Perspectives (2025). https://doi.org/10.2139/ssrn.5218304
30. Greenberg, J.: The social side of fairness: Interpersonal and informational classes of organizational justice. Justice in the Workplace: Approaching Fairness in Human Resource Management. pp. 79–106 (1993)
31. Colquitt, J.A.: Organizational justice. In: Kozlowski, S.W.J. (ed.) The Oxford Handbook of Organizational Psychology, pp. 526–547. Oxford University Press (2012)
32. Decker, M., Wegner, L., Leicht-Scholten, C.: Procedural fairness in algorithmic decision-making: the role of public engagement. Ethics Inf. Technol. **27**(1) (2025). https://doi.org/10.1007/s10676-024-09811-4
33. Lee, M.K., Jain, A., Cha, H.J., Ojha, S., Kusbit, D.: Procedural Justice in Algorithmic Fairness. Proc. ACM Hum. Comput Interact (CSCW). **3**:1–26 (2019). https://doi.org/10.1145/3359284
34. Grgic-Hlaca, N., Zafar, M.B., Gummadi, K.P., Weller, A.: The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In Symposium on Machine Learning and the Law, Barcelona, Spain (2016)
35. Heilinger, J.-C.: The Distributive Demands of Relational Egalitarianism. Ethic Theory Moral. Prac. **27**(4), 619–634 (2024). https://doi.org/10.1007/s10677-024-10460-1
36. Sen, A.: Equality of What? (1979)
37. Dworkin, R.: What is Equality? Part 2: Equality of Resources. Philos. Public Affairs. **10**(4), 283–345 (1981). https://doi.org/10.4324/9781315199795-7
38. Arneson, R.: Equality and equal opportunity for welfare. Philos. Stud. **56**(1), 77–93 (1989). https://doi.org/10.1007/bf00646210
39. Nussbaum, M.C.: Capabilities as fundamental entitlements. Fem. Econ. **9**, 33–59 (2003)
40. Temkin, L.S.: Inequality. Oxford University Press, New York (1993)
41. Cohen, G.A.: On the Currency of Egalitarian Justice. Ethics. **99**(4), 906–944 (1989). https://doi.org/10.1086/293126
42. Voigt, K.: The Harshness Objection: Is Luck Egalitarianism Too Harsh on the Victims of Option Luck? Ethic Theory Moral. Prac. **10**(4), 389–407 (2007). https://doi.org/10.1007/s10677-006-9060-4

43. Scheffler, S.: The Practice of Equality. In: Fourie, C., Schuppert, F., Wallimann-Helmer, I. (eds.) Social Equality. Oxford University Press (2015)
44. Young, I.M.: Responsibility for justice. Oxford University Press, New York, Oxford (2011)
45. Young, I.M.: Responsibility and global justice: A social connection model. Social Philos. policy. **23**(1), 102–130 (2006)
46. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A Survey on Bias and Fairness in Machine Learning (2019). https://doi.org/10.48550/arXiv.1908.09635
47. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference pp. 214–226 (2012). https://doi.org/10.1145/2090236.2090255
48. Heidari, H., Loi, M., Gummadi, K.P., Krause, A.: A moral framework for understanding fair ML through economic models of equality of opportunity. In Proceedings of the conference on fairness, accountability, and transparency. pp. 181–190 (2019). https://doi.org/10.1145/3287560.3287584
49. Franke, U.: Rawls's Original Position and Algorithmic Fairness. Philos. Technol. **34**(4), 1803–1817 (2021). https://doi.org/10.1007/s13347-021-00488-x
50. Edward, N., Zalta (eds.): The Stanford Encyclopedia of Philosophy, 2017th edn. Metaphysics Research Lab, Stanford University (2017)
51. Kitchenham, B.: Procedures for performing systematic reviews, Keele, UK, Keele University, vol. 33, pp. 1–26, (2004). [Online] Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=29890a936639862f45cb9a987dd599dce9759bf5
52. Snyder, H.: Literature review as a research methodology: an overview and guidelines. J. Bus. Res. **104**, 333–339 (2019). https://doi.org/10.1016/j.jbusres.2019.07.039
53. Nath, R.: Relational egalitarianism, Philosophy Compass, vol. 15, no. 7 (2020). https://doi.org/10.1111/phc3.12686
54. Collins, P.H., Bilge, S.: Intersectionality. Cambridge, UK: Polity Press, (2016). [Online] Available: https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=4698012
55. Klein, L., D'Ignazio, C.: Data Feminism for AI In The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), pp. 100–112 (2024). https://doi.org/10.1145/3630106.3658543
56. Aizenberg, E., van den Hoven, J.: Designing for human rights in AI. Big Data Soc. **7**(2) (2020). https://doi.org/10.1177/2053951720949566
57. Ajmani, L., Stapleton, L., Houtti, M., Chancellor, S. (eds.): Data Agency Theory: A Precise Theory of Justice for AI Applications (2024). https://doi.org/10.1145/3630106.3658930
58. Andrus, M., Villeneuve, S.: Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness In 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, pp. 1709–1721 (2022). https://doi.org/10.1145/3531146.3533226
59. Birhane, A.: Algorithmic injustice: a relational ethics approach, Patterns (New York, N.Y.), vol. 2, no. 2, (2021). https://doi.org/10.1016/j.patter.2021.100205
60. Birhane, A., et al.: Power to the People? Opportunities and Challenges for Participatory AI In Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22), Arlington, VA, USA. (2022). [Online] Available: https://dl.acm.org/doi/pdf/10.1145/3551624.3555290. Accessed 22 May 2023.
61. Braun, M., Hummel, P.: Data justice and data solidarity Patterns (New York, N.Y.), vol. 3, no. 3, pp. 100427 (2022). https://doi.org/10.1016/j.patter.2021.100427
62. Burrell, J.: Automated decision-making as domination. First Monday. (2024). https://doi.org/10.5210/fm.v29i4.13630
63. Burrell, J., Fourcade, M.: The Society of Algorithms. Annu. Rev. Sociol. **47**(1), 213–237 (2021). https://doi.org/10.1146/annurev-soc-090820-020800
64. Cinnamon, J.: Social Injustice in Surveillance Capitalism, Surveillance & Society, vol. 15, no. 5, pp. 609–625 (2017). https://doi.org/10.24908/ss.v15i5.6433
65. Davis, J.L., Williams, A., Yang, M.W.: Algorithmic reparation. Big Data Soc. **8**(2) (2021). https://doi.org/10.1177/20539517211044808
66. Gangadharan, S.P., Niklas, J.: Decentering technology in discourse on discrimination Information, Communication & Society, vol. 22, no. 7, pp. 882–899 (2019). https://doi.org/10.1080/1369118X.2019.1593484
67. Gwagwa, A., Kazim, E., Hilliard, A.: The role of the African value of Ubuntu in global AI inclusion discourse: a normative ethics perspective. Patterns (New York, N.Y.). **3**(4), 1–7 (2022). https://doi.org/10.1016/j.patter.2022.100462
68. Heilinger, J.-C.: The Ethics of AI Ethics. A Constructive Critique. Philos. Technol. **35**(3) (2022). https://doi.org/10.1007/s13347-022-00557-9
69. Himmelreich, J.: Against 'Democratizing AI' AI & Soc, vol. 38, no. 4, pp. 1333–1346 (2023). https://doi.org/10.1007/s00146-021-01357-z
70. Huang, L.T.-L., Chen, H.-Y., Lin, Y.-T., Huang, T.-R., Hun, T.-W.: Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy. Feminist Philos. Q. **8**(3/4) (2022). https://doi.org/10.5206/fpq/2022.3/4.14347
71. Kassam, A., Marino, P.: Algorithmic Racial Discrimination: A Social Impact Approach. Feminist Philos. Q. **8** (2022). https://doi.org/10.5206/fpq/2022.3/4.14275 3/4
72. Kong, Y.: Are 'Intersectionally Fair' AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, pp. 485–494 (2022). https://doi.org/10.1145/3531146.3533114
73. Krupiy, T.: A vulnerability analysis: Theorising the impact of artificial intelligence decision-making processes on individuals, society and human diversity from a social justice perspective. Comput. Law Secur. Rev. **38**, 105429 (2020). https://doi.org/10.1016/j.clsr.2020.105429
74. Leavy, S., Siapera, E., O'Sullivan, B.: Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race. In: AIES '21, pp. 695–703. Virtual Event, USA (2021). https://doi.org/10.1145/3461702.3462598
75. Li, T., Iacobelli, F.: Purposeful AI In: Computer Supported Cooperative Work and Social Computing, Minneapolis, MN, USA, pp. 563–565 (2023). https://doi.org/10.1145/3584931.3606954
76. Lin, T.-A., Chen, P.-H.C.: Artificial Intelligence in a Structurally Unjust Society. Feminist Philos. Q. **8** (2022). https://doi.org/10.5206/fpq/2022.3/4.14191 3/4
77. Lopez, P.: More than the Sum of its Parts: Susceptibility to Algorithmic Disadvantage as a Conceptual Framework In The 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil, pp. 909–919 (2024). https://doi.org/10.1145/3630106.3658944
78. Lu, C., Kay, J., McKee, K.R.: Subverting machines, fluctuating identities: Re-learning human categorization In FAccT '22, pp. 1005–1015 (2022). https://doi.org/10.1145/3531146.3533161
79. McFadden, Z., Alvarez, L.: Performative Ethics From Within the Ivory Tower: How CS Practitioners Uphold Systems of Oppression. J. Artif. Intell. Res. (2024). https://doi.org/10.1613/jair.1.15423
80. Medrado, A., Verdegem, P.: Participatory action research in critical data studies: interrogating AI from a South–North approach. Big Data Soc. **11**(1) (2024). https://doi.org/10.1177/20539517241235869

81. Miragoli, M.: Conformism, ignorance & injustice: AI as a tool of epistemic oppression. Episteme. 1–19 (2024). https://doi.org/10.1017/epi.2024.11
82. Naudts, L.: The Digital Faces of Oppression and Domination: A Relational and Egalitarian Perspective on the Data-driven Society and its Regulation. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil, pp. 701–712 (2024). https://doi.org/10.1145/3630106.3658934
83. Park, J.S., Karahalios, K., Salehi, N., Eslami, M.: Power Dynamics and Value Conflicts in Designing and Maintaining Socio-Technical Algorithmic Processes. Proc. ACM Hum. Comput Interact (CSCW1). **6**, 1–21 (2022). https://doi.org/10.1145/3512957
84. Rafanelli, L.M.: Justice, injustice, and artificial intelligence: Lessons from political theory and philosophy. Big Data Soc. (2022). https://doi.org/10.1177/20539517221080676
85. Sloane, M., Moss, E., Awomolo, O., Forlano, L.: Participation Is not a Design Fix for Machine Learning. In EAAMO '22: Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 1–6 (2022). https://doi.org/10.1145/3551624.3555285
86. Stone, J., Mittelstadt, B.: Legitimate Power, Illegitimate Automation: The problem of ignoring legitimacy in automated decision systems In: The 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil (2025). https://doi.org/10.1145/3725858
87. Tacheva, J., Ramasubramanian, S., Empire, A.I.: Unraveling the interlocking systems of oppression in generative AI's global order (2023). https://doi.org/10.1177/20539517231219241
88. Tacheva, Z.: Tracking a critical look at the critical turn in data science: from 'data feminism' to transnational feminist data science. Big Data Soc. **9**(2) (2022). https://doi.org/10.1177/20539517221112901
89. Tomasev, N., McKee, K.R., Kay, J., Mohamed, S.: Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event USA, pp. 254–265 (2021). https://doi.org/10.1145/3461702.3462540
90. Vecchione, B., Levy, K., Barocas, S.: Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies In Equity and Access in Algorithms, Mechanisms, and Optimization, NY, USA, pp. 1–9 (2021). https://doi.org/10.1145/3465416.3483294
91. Wong, P.-H., Fairness, D.A.: Philos. Technol. **33**(2), 225–244 (2020). https://doi.org/10.1007/s13347-019-00355-w
92. Zajko, M.: Conservative AI and social inequality: conceptualizing alternatives to bias through social theory. AI Soc. **36**(3), 1047–1056 (2021). https://doi.org/10.1007/s00146-021-01153-9
93. Zhang, A., Hosoi, A.: Structural Interventions and the Dynamics of Inequality In: The 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil, pp. 1014–1030 (2024). https://doi.org/10.1145/3630106.3658952
94. Zhang, M.: Affirmative Algorithms: Relational Equality as Algorithmic Fairness In: FAccT 2022: 2022 5th ACM Conference on Fairness, Accountability, and Transparency: June 21–24, Seoul, South Korea, Seoul Republic of Korea, pp. 495–507 (2022). https://doi.org/10.1145/3531146.3533115
95. Zimmermann, A., Lee-Stronach, C.: Proceed with Caution. Can. J. Philos. **52**(1), 6–25 (2022). https://doi.org/10.1017/can.2021.17
96. Ahmed, S.K., et al.: Using thematic analysis in qualitative research. J. Med. Surg. Public. Health. **6**, 100198 (2025). https://doi.org/10.1016/j.glmedi.2025.100198
97. Clarke, V., Braun, V.: Teaching thematic analysis: overcoming challenges and developing strategies for effective learning. Psychologist. **26**, 2 (2013)
98. Braun, V., Clarke, V.: Supporting best practice in reflexive thematic analysis reporting in Palliative Medicine: a review of published research and introduction to the Reflexive Thematic Analysis Reporting Guidelines (RTARG). Palliat. Med. **38**(6), 608–616 (2024). https://doi.org/10.1177/02692163241234800
99. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative Res. Psychol. **3**(2), 77–101 (2006). https://doi.org/10.1191/1478088706qp063oa
100. Greene, D., Hoffmann, A.L., Stark, L.: Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning In: Proceedings of the Annual Hawaii International Conference on System Sciences (2019)
101. Costanza-Chock, S.: Design Justice: Community-Led Practices to Build the Worlds We Need. The MIT Press, Cambridge, MA (2020)
102. König, P.D., Wenzelburger, G.: The legitimacy gap of algorithmic decision-making in the public sector: Why it arises and how to address it. Technol. Soc. **67**, 101688 (2021). https://doi.org/10.1016/j.techsoc.2021.101688
103. Santoni de Sio, F., Mecacci, G.: Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. Philos. Technol. **34**(4), 1057–1084 (2021). https://doi.org/10.1007/s13347-021-00450-x
104. Nyholm, S.: Responsibility Gaps, Value Alignment, and Meaningful Human Control over Artificial Intelligence. In: Placani, A., Broadhead, S. (eds.) in Risk and Responsibility in Context, pp. 191–213. Routledge, New York (2023). https://doi.org/10.4324/9781003276029-14
105. Königs, P.: Artificial intelligence and responsibility gaps: what is the problem? In: Ethics Inf Technol, vol. 24, no. 3, pp. 1–11 (2022). https://doi.org/10.1007/s10676-022-09643-0
106. Johnson, M.: Embodied Knowledge Curriculum Inquiry. vol. 19, no. 4, pp. 361–377 (1989). https://doi.org/10.1080/03626784.1989.11075338
107. Howells, J.: Tacit knowledge Technology Analysis & Strategic Management. vol. 8, no. 2, pp. 91–106 (1996). https://doi.org/10.1080/09537329608524237
108. Sangiovanni, A., Viehoff, J.: Solidarity in Social and Political Philosophy The Stanford Encyclopedia of Philosophy, Winter 2023 Edition. Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/win2023/entries/solidarity/ (2023)
109. Maas, J., Moreno, A., Inglés, Beyond Participatory, A.I.: AIES. **7**, 932–942 (2024). https://doi.org/10.1609/aies.v7i1.31693
110. Santoni, F., de Sio: Human Freedom in the Age of AI. Routledge, New York (2024)
111. Nost, E., Colven, E.: Earth for AI: A Political Ecology of Data-Driven Climate Initiatives. Geoforum. **130**, 23–34 (2022). https://doi.org/10.1016/j.geoforum.2022.01.016
112. Adams, R.: Can artificial intelligence be decolonized? Interdisc. Sci. Rev. **46**, 1–2 (2021). https://doi.org/10.1080/03080188.2020.1840225
113. Muldoon, J., Wu, B.A.: Artificial Intelligence in the Colonial Matrix of Power. Philos. Technol. **36**(4) (2023). https://doi.org/10.1007/s13347-023-00687-8
114. French, J.R., Raven, B.: The bases of social power. Stud. social power, pp. 150–167 (1959)
115. Pfeffer, J., Salancik, G.R.: The external control of organizations: A resource dependence perspective. Stanford, Calif. (2003)
116. Foucault, M.: Discipline and Punish: The Birth of the Prison, 2nd edn. Vintage Books, New York (1995)
117. Arendt, H., Allen, D.S., Canovan, M.: The human condition. The University of Chicago Press, Chicago, London (2018)
118. Ovalle, A., Subramonian, A., Gautam, V., Gee, G., Chang, K.-W.: Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. In: Proceedings

of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, Montreal, QC, Canada, pp. 496–511 (2023). https://doi.org/10.1145/3600211.3604705

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.